THE CONVOLUTION OF THE NORMAL AND LOGNORMAL DISTRIBUTIONS

Douglas M. Hawkins
Department of Applied Statistics
Technical Report No. 520
School of Statistics
University of Minnesota
November 7, 1988

THE CONVOLUTION OF THE NORMAL AND LOGNORMAL DISTRIBUTIONS

Douglas M. Hawkins
Department of Applied Statistics
School of Statistics
University of Minnesota
St Paul
MN 55108

Abstract

Let Z be a three parameter lognormal variate, Y a normal with zero mean and define X = Z+Y. The marginal distribution of X is then the convolution of the lognormal with the normal - a distribution we will abbreviate to LNN. Expressions for the density and distribution function of the LNN are given, and its properties sketched. Maximum likelihood, moment and modified moment estimators of the parameters of the LNN are given. Some applications require for calibration the conditional distribution of X given Z. This is derived, along with the calibration curve E(X|Z). An application to X-ray fluorescence counting of in situ gold grades is discussed.

Key Words: estimation, calibration, conditional distribution.

Author's footnote

Douglas M. Hawkins is Professor, Department of Applied Statistics, University of Minnesota. The initial theoretical developments in this paper were made under the aegis of the National Research Institute for Mathematical Sciences, Council for Scientific and Industrial Research, Pretoria, South Africa. The computations were carried out under Research Support Program RPS 1027 of the International Business Machines Corporation. The author gratefully acknowledges IBM's computing support; the permission of the Chamber of Mines Research Organization to make use of data and confidential reports by himself and Professor H S Sichel; and the helpful discussions about numerical analysis held with Dr D P Laurie.

1. Introduction

The lognormal distribution has found application in many problem areas (Aitchison and Brown 1957) in a variety of fields ranging from biology to economics. It is particularly useful in mining, as the De Wijzian model of ore deposits (Matheron 1960) suggests that it may apply to any mineral present in low concentrations, and this has been verified experimentally for oil and many minerals (Krige 1978, Harbaugh and Ducastaing 1981). The three parameter lognormal was introduced to geology by Krige (1960), for modelling gold and uranium grades and is now widely regarded as the 'natural' parametric model for low-concentration deposits.

It has been the standard practice when studying lognormal data to ignore any measurement error there might be in observing the lognormal variable. For many applications, this assumption is justified by the high precision of measurement relative to the true variability. Recently however there has been a growing interest in the mining community in measuring lognormally distributed quantities with instruments which give quick and inexpensive readings that have substantial random measurement errors. One example of this is the measurement of uranium grades using scintillation counters. Another is a technique in which ore is irradiated with high energy electromagnetic sources (Rolle, 1979) causing each element in the ore (including those of economic interest) to fluoresce at specific wavelengths. The counts in the valuable element's 'channel' provide an estimate of the concentration (or 'grade') of the element, but this has first to be corrected for background radiation, which is done by subtracting the counts obtained in a 'background' channel. This necessary background correction gives rise to a measurement random error with a variance much too large to ignore - for example these procedures often give negative estimates of grade. It was this instrument and its properties that gave rise to the work discussed in this paper. The same

statistical framework is plausible also in other contexts where the lognormal distribution arises. For example the lognormal is a natural model for low concentration pollutants, and where for reasons of economy these are estimated using indirect assay methods, the same measurement framework arises.

Several statistical problems arise within this framework. One is the estimation of the parameters of the underlying lognormal distribution using the LNN readings. Another is the calibration problem - conditional on the LNN reading, what is the distribution of the true grade; and derived from this, the calibration curve and its standard error. These problems will be addressed below.

2. Derivation of the LNN distribution

Let us suppose that the underlying quantity of interest Z follows a three-parameter lognormal distribution (3PLND):-

$$\ell n(Z+\alpha) \sim N(\xi, \sigma^2).$$

We assume that X is a measurement of Z with a normally distributed random measurement error U -

$$X = Z + U, \text{ where } U \sim N(0, \tau^2).$$

so that conditionally on Z,

$$X \sim N(Z, \tau^2),$$

and the measurement X of Z is conditionally unbiased with constant variance τ^2. Then X follows the lognormal-normal convolution (LNN) distribution.

In the development following we will assume τ known. This is the case in the problem motivating the study (where τ can be computed from the theory of Poisson counting statistics), and pertains in many potential applications of

the LNN. Where τ is not known, it would be best measured (if possible) by repeat measures X of the same Z, with its estimation directly from the X readings a poor third choice. At the appropriate point however, we will indicate the minor adaptations needed to estimate τ along with the other parameters.

Define the two-parameter lognormal quantity (2PLND) $Y = Z + \alpha$. Then the joint distribution of X and Y is easily seen to be

$$f(x,y) = f(y)\, f(x|y)$$

$$= \frac{1}{2\pi\sigma\tau y} \exp -\frac{1}{2}\left[\frac{(\ell n\, y - \xi)^2}{\sigma^2} + \frac{(x - y + \alpha)^2}{\tau^2}\right].$$

The marginal density and cumulative distribution function of X are then

$$g(x|\alpha,\tau,\xi,\sigma) = \int_{y=0}^{\infty} f(x,y)\, dy$$

$$= \int_{y=0}^{\infty} \frac{1}{2\pi\sigma\tau y} \exp -\frac{1}{2}\left[\frac{(\ell n\, y - \xi)^2}{\sigma^2} + \frac{(x - y + \alpha)^2}{\tau^2}\right]$$

$$G(x|\alpha,\tau,\xi,\sigma) = \int_{y=0}^{\infty} f(y)\, \Pr[X \le x | Y = y)]\, dy$$

$$= \int_{y=0}^{\infty} \bar{\Phi}\left[\frac{x - y + \alpha}{\tau}\right] \frac{1}{\sqrt{(2\pi)}\sigma y} \exp -\frac{1}{2}\left[\frac{(\ln y - \xi)^2}{\sigma^2}\right] dy \qquad \ldots(1)$$

where $\bar{\Phi}(.)$ denotes the standard normal distribution function.

Alternative and computaitonally more convenient expressions may be obtained by making the transformation $t = \ell n\, y$. This gives:

$$g(x|\alpha,\tau,\xi,\sigma) = \int_{t=-\infty}^{\infty} \frac{1}{2\pi\sigma\tau} \exp -\frac{1}{2}\left[\frac{(t-\xi)^2}{\sigma^2} + \frac{(x - e^t + \alpha)^2}{\tau^2} \right] dt$$

$$G(x|\alpha,\tau,\xi,\sigma) = \int_{t=-\infty}^{\infty} \bar{\Phi}\left[\frac{x - e^t + \alpha}{\tau}\right] \frac{1}{\sqrt{(2\pi)}\sigma} \exp -\frac{1}{2}\left[\frac{(t-\xi)^2}{\sigma^2} \right] dt \qquad ...(2)$$

Neither representation gives a standard well-known integrals, and so they must be evaluated numerically. The details of this are discussed in Section 5.

The canonical form of the LNN

The parameters α and τ are essentially location and scale parameters for the LNN. Specifically, if write $X' = aX + b$, then it is easily seen that X' has the density $g(x'|a\alpha - b, a\tau, \xi + \ell n\ a, \sigma)$.

In particular, $X' = (X + \alpha)/\tau$ has the 'canonical' density of the LNN $g(x'|0,1,\xi-\ell n\ \tau,\sigma)$. From this, we obtain

$$g(x|\alpha,\tau,\xi,\sigma) = g(\{x + \alpha\}/\tau|\alpha,\tau,\xi-\ell n\ \tau,\sigma)/\tau$$
$$G(x|\alpha,\tau,\xi,\sigma) = G(\{x + \alpha\}/\tau|\alpha,\tau,\xi,\sigma)$$

so that only the standard canonical form $g(x|0,1,\xi,\sigma)$ need be studied. The density and cumulative distribution function of this canonical form are

$$g(x|0,1,\xi,\sigma) = \int_{t=-\infty}^{\infty} \frac{1}{2\pi\sigma} \exp -\frac{1}{2}\left[\frac{(t-\xi)^2}{\sigma^2} + (x-e^t)^2 \right] dt$$

$$G(x|0,1,\xi,\sigma) = \int_{t=-\infty}^{\infty} \bar{\Phi}\left[x-e^t\right] \frac{1}{\sqrt{(2\pi)}\sigma} \exp -\frac{1}{2}\left[\frac{(t-\xi)^2}{\sigma^2} \right] dt \qquad ...(3)$$

3 Properties of the LNN (i) Moments

Certain properties of the marginal distribution of X may be inferred directly from the definition - for example the moments. Since $X = Z+U$, where $U \sim N(0,\tau^2)$ and is independent of Z,

$$E(X^k) = E(Z + U)^k$$
$$= \sum_{j=0}^{k} \binom{k}{j} E(Z^j)\, E(U^{k-j}) \qquad \ldots(4)$$

This expression is easily evaluated using the known moments of lognormal and of normal variates. In particular, if we transform to the canonical variate $X' = (x + \alpha)/\tau$, the first four moments are given by

$$E(X') = \exp(\xi' + \tfrac{1}{2}\sigma^2)$$
$$E(X'^2) = 1 + \exp(2\xi' + 2\sigma^2)$$
$$E(X'^3) = \exp(\xi' + \tfrac{1}{2}\sigma^2)\,\{3 + \exp(2\xi' + 4\sigma^2)\}$$
$$E(X'^4) = 3 + 6\exp(2\xi' + 2\sigma^2) + \exp(4\xi' + 8\sigma^2) \qquad \ldots(5)$$

Note that the sequence of moments of X increases more rapidly than that of the underlying lognormally distributed Z, and that the distribution of X, like the 3PLND, is too heavy-tailed to be determined by its moments.

(ii) Asymptotics

It is evident from the definition of X that its extreme right tail is that of the underlying 3PLND variable Z. Specifically, as $x \to \infty$,

$$G(x|\alpha,\tau,\xi,\sigma) \sim \bar{\Phi}\{(\ell n\, x+\alpha-\xi)/\sigma\}.$$

In particular, on lognormal probability paper, the LNN data will give a

straight line out to the far right. Provided this linear segment is visually clear and long enough, this provides a quick graphic estimate of the underlying lognormal parameters.

The left tail behavior is less simple. Expression 3 shows that

$$G(x|0,1,\xi,\sigma) < \bar{\Phi}(x)$$

with asymptotic equality if ξ is large negative (under which circumstances the lognormal part of X is negligible). If σ is near zero, then G approximates $\bar{\Phi}[x-\exp(\xi)]$. In the more interesting case that ξ is positive and σ appreciable, then the left tail beahvior is not particularly straightforward.

Also interesting is the behavior for extreme values of the parameters. It is well known that as $\sigma \to 0$ the lognormal approaches the normal (see for example Klimko et al 1975 or Kotz, 1973). Any normal distribution $N(\mu,\tau^2)$ say can be obtained as the degenerate limiting case of the 3PLND letting $\alpha \to \infty$ and $\sigma \to 0$ while maintaining the moment equations

$$\alpha + \exp(\xi+\tfrac{1}{2}\sigma^2) = \mu, \quad \exp(2\xi+2\sigma^2)[\exp(\sigma^2)-1] = \tau^2.$$

SInce the distribution of Z approaches the normal under certain circumstances, then so also must that of the LNN.

This fact that the normal distribution is a special but degenerate limiting case of the LNN has important implications for parameter estimation, as we shall see below.

4. Inference on the parameters

Apart from the parameter τ, which is assumed known, the distribution g has the three parameters α, ξ and σ whose values must be established before any computations can be made. Let us suppose therefore that we have a sample size n, $X_1,\ldots,X_n$ available for this estimation.

(a) Methods of Moments

The key to the moment estimators lies in the fact that the first and third central moments of X are identical to those of Z, while the variance of X exceeds that of Z by τ^2. Thus, defining

$$\bar{X} = \sum_{i=1}^{n} X_i / n$$

$$s^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2/n - \tau^2$$

$$m_3 = \sum_{i=1}^{n} (X_i - \bar{X})^2/ n$$

$$b_1 = m_3^2 / s^3$$

we obtain estimates of the first three moments of the underlying 3PLND variable Z, and are able to apply the moment equations of the 3PLND, whose solution (Johnson and Kotz 1970 p. 124) is:

$$D = \{(1 + \tfrac{1}{2}b_1)^2 - 1\}^{\frac{1}{2}}$$

$$\omega = \sqrt[3]{(1 + \tfrac{1}{2}b_1 + D)} + \sqrt[3]{(1 + \tfrac{1}{2}b_1 - D)} - 1$$

$$\hat{\sigma}^2 = \log \omega$$

$$\hat{\xi} = \tfrac{1}{2} \ell n \left[s / \{\omega(\omega - 1)\}\right]$$

$$\hat{\alpha} = \exp(\hat{\xi} + \tfrac{1}{2}\hat{\sigma}^2) - \bar{X} \qquad \ldots(6)$$

The attraction of these moment estimators of the parameters is that they are in closed form and are very easily computed. Their major disadvantage is that since the LNN distribution is very heavy tailed, they can be expected to be quite inefficient.

There is a noteworthy difference between the method of moment estimators of the LNN and of the 3PLND - s^2 is not the variance of the data X, but the

variance less the counting variance τ^2. It is thus possible (unlike the 3PLND case) for s^2 to be negative when the estimation fails. This is an indication that ξ and/or σ are small. As in the 3PLND case, solution of these equations requires that the sample skewness b_1 be strictly positive. The skewness is very strongly related to a degeneracy diagnostic defined below, and a negative value of b_1 is a strong indication that the data conform better to the normal than the LNN distribution, and that the best fit will be obtained by degenerate parameter values.

(b) Modified method of moments estimation

An unattractive feature of the method of moments estimator is its use of the sample skewness, a statistic which is likely to be poorly estimated in samples of moderate size because of the very heavy tail of the LNN distribution. To ameliorate the corresponding problem in the 3PLND distribution Cohen and Whitten (1980) proposed modified method of moments estimators, in which the equation based on skewness is replaced by some other equation. Particularly simple and effective is the system obtained by matching a sample quantile instead of the sample skewness.

This suggests the adaptation of the same idea to the LNN distribution. Here however the use of quantiles is not as simple as in the 3PLND case, since the LNN quantiles (unlike the 3PLND) are not simple functions of the parameters. There is however one quantile of the LNN which is fairly close to a simple fixed function of the parameters. The median turns out to be quite close to $\exp(\xi)$ in a wide range of parameter values. (This is exactly true for the underlying 3PLND). Table 1 shows the ratio of the true median to $\exp(\xi)$ for a series of different values of ξ and σ in the canonical parametrization. While the table confirms that the true median is in general not equal to $\exp(\xi)$, the discrepancy between the two is marked only when ξ is negative - a situation in which estimation by any means is difficult because the lognormal signal tends to be drowned in the normal measurement noise.

This suggests the use of the following procedure - find Q_2, the median of the sample, and then solve

$$\bar{X} = -\alpha + e^{\xi}\sqrt{\omega}$$
$$s^2 = e^{2\xi}\omega(\omega-1)$$
$$Q_2 = -\alpha + e^{\xi}. \qquad \ldots(7)$$

A little manipulation of these equations yields a cubic in α:-

$$\alpha^3 + b_2\alpha^2 + b_1\alpha + b_0 = 0, \text{ where}$$
$$b_2 = \tfrac{1}{2}(Q_2 + 5\bar{X} - R)$$
$$b_1 = (\bar{X}Q_2 + 2\bar{X}^2 - RQ_2)$$
$$b_0 = \tfrac{1}{2}(\bar{X}^2Q_2 + \bar{X}^3 - RQ_2^2)$$
$$\text{where } R = s^2/(\bar{X}-Q_2). \qquad \ldots(8)$$

Feasibility requires that $\bar{X}>Q_2$. Provided this is the case, we may solve the cubic for α, getting either one or three real roots. Given α, the first two equations of (7) yield solutions for e^{ξ} and $\sqrt{\omega}$. It is possible for these estimates to be negative (this occurs when the estimated α is smaller than $-Q_2$), and then the method fails.

If the method does not fail, and the estimated ξ is found to be negative so that the approximation of the true median by $\exp(\xi)$ is suspect, then the estimated ξ and σ may be entered in Table 1 to get a correction factor. Multiplying Q_2 by this factor and solving the equations again will remove most of the effect of the approximation of the median by $\exp(\xi)$, though it is questionable whether this correction is really warranted in practice.

(c) Method of Maximum Likelihood

In the defining equation of $g(x|\alpha,\tau,\xi,\sigma)$, let

$$A_i(t) = \frac{1}{2\pi\sigma\tau} \exp -\frac{1}{2}\left[\frac{(t-\xi)^2}{\sigma^2} + \frac{(X_i - e^t + \alpha)^2}{\tau^2} \right]$$

Using the Macsyma symbolic computation system the first and second derivatives of $g(X_i|\alpha,\tau,\xi,\sigma)$ were found to be:

$$\frac{\partial g}{\partial \alpha} = \int_{t=-\infty}^{\infty} (e^t - X_i - \alpha)\, A_i(t)\, dt \,/\, \tau^2$$

$$\frac{\partial g}{\partial \xi} = \int_{t=-\infty}^{\infty} (t - \xi)\, A_i(t)\, dt \,/\, \sigma^2$$

$$\frac{\partial g}{\partial \sigma^2} = \int_{t=-\infty}^{\infty} \{(t-\xi)^2 - \sigma^2\}\, A_i(t)\, dt \,/\, 2\sigma^4$$

$$\frac{\partial g}{\partial \tau^2} = \int_{t=-\infty}^{\infty} \{(e^t - X_i - \alpha) - \tau^2\} A_i(t)\, dt \,/\, 2\tau^2$$

$$\frac{\partial^2 g}{\partial \alpha^2} = \int_{t=-\infty}^{\infty} \{(e^t - X_i - \alpha)^2 - \tau^2\}\, A_i(t)\, dt \,/\, 2\tau^4$$

$$\frac{\partial^2 g}{\partial \alpha \partial \xi} = \int_{t=-\infty}^{\infty} (t - \xi)(e^t - X_i - \alpha)\, A_i(t)\, dt \,/\, \sigma^2\tau^2$$

$$\frac{\partial^2 g}{\partial \xi^2} = \int_{t=-\infty}^{\infty} \{(t - \xi)^2 - \sigma^2\}\, A_i(t)\, dt \,/\, \sigma^4$$

$$\frac{\partial^2 g}{\partial \alpha \partial \sigma^2} = \int_{t=-\infty}^{\infty} (e^t - X_i - \alpha)\{(t-\xi)^2 - \sigma^2\}\, A_i(t)\, dt \,/\, 2\sigma^4\tau^2$$

$$\frac{\partial^2 g}{\partial \xi \partial \sigma^2} = \int_{t=-\infty}^{\infty} (t - \xi)\{(t - \xi)^2 - 3\sigma^2\}\, A_i(t)\, dt \,/\, 2\sigma^6$$

$$\frac{\partial^2 g}{\partial (\sigma^2)^2} = \int_{t=-\infty}^{\infty} \{(t - \xi)^4 - 6\sigma^2(t - \xi)^2 + 3\sigma^4\}\, A_i(t)\, dt \,/\, 4\sigma^8$$

$$\frac{\partial^2 g}{\partial \alpha \partial \tau^2} = \int_{t=-\infty}^{\infty} (e^t - X_i - \alpha)\{(e^t - X_i - \alpha)^2 - 3\tau^2\}\, A_i(t)\, dt \,/\, 2\tau^6$$

$$\frac{\partial^2 g}{\partial\xi\partial\tau^2} = \int_{t=-\infty}^{\infty} (t-\xi)\{(e^t-X_i-\alpha)^2 - \tau^2\}\, A_i(t)\, dt \,/\, 2\tau^4\sigma^2$$

$$\frac{\partial^2 g}{\partial\sigma^2\partial\tau^2} = \int_{t=-\infty}^{\infty} \{(t-\xi) - \sigma^2\}\{(e^t-X_i-\alpha)^2 - \tau^2\}\, A_i(t)\, dt \,/\, 4\tau^4\sigma^4$$

$$\frac{\partial^2 g}{\partial(\tau^2)^2} = \int_{t=-\infty}^{\infty} \{(e^t-X_i-\alpha)^4 - 6\tau^2(e^t-X_i-\alpha)^2 + 3\tau^4\}\, A_i(t)\, dt \,/\, 4\tau^8 \qquad \ldots(9)$$

Although in our applications we will assume τ known, for the sake of completeness the list above includes the derivatives of g with respect to τ as well as the other three parameters.

From these expressions, the derivatives of the log likelihood function follow at once, and these in turn may be used to compute maximum likelihood estimators (MLE's) of α, ξ and σ iteratively. If a second order method is used to maximize the log likelihood, then the observed Fisher information matrix is obtained as a by-product, and can be used to obtain approximate (asymptotic) standard errors for the parameter estimates.

Unlike the situation with the 3PLND, the likelihood has no singularities and is infinitely differentiable with respect to all of its parameters. It is of interest to note that one of the methods of avoiding the singularity problem of the 3PLND (see for example Griffiths 1980) consists of a recognition that the observed values always invole some error, even if it be only recording to a finite number of decimals and recognition of this fact leads to a modified likelihood equation in which the singularity not present.

The LNN shares with the 3PLND another form of degeneracy however - that in which the data appear to conform better to the normal than the LNN distribution. When data of this type are subjected to maximum likelihood estimation, the successive estimates diverge, with $\alpha\to\infty$, $\sigma\to 0$, and ξ 'taking up the slack', implicitly fitting the normal distribution as the degenerate asymptotic LNN. Klimko et al (1975) describe this problem as tending to arise

in 3PLND data when the sample has a negative skewness, but a more precise analysis would be along the following lines:

Take a 3PLND sample - for example the underlying Z_i of this paper. Assume $\hat{\alpha}$ given and write

$$S_k = \sum_{i=1}^{n} [\ell n\, Z_i + \alpha)^k] \,/n.$$

Then the 3PLND likelihood maximized over ξ and σ may be written

$$2\, \ell n\, L = -n - 2S_1 - \ell n\, [S_2 - S_1^2]. \qquad \ldots(10)$$

A necessary condition for the MLE to be degenerate is that this quantity is increasing as $\alpha \to \infty$. Differentiating with respect to α, we find that this is true if:

$$\sum_{i=1}^{n} \frac{1}{Z_i+\alpha} [\ell n\, (Z_i+\alpha) - S_1 + S_2 - S_1^2] < 0 \text{ for large } \alpha. \qquad \ldots(11)$$

A diagnostic for degeneracy of the 3PLND is therefore to compute (11) for a sufficiently large value of α. If its value is negative, this suggests the ML solution would be degenerate; if it is positive, it proves that the solution is not degenerate.

Klimko et al use the sample skewness as an indicator of likely degeneracy, but show that the skewness is not always a reliable indicator of degeneracy. Evaluating the skewness and (11) in a variety of test cases has shown that while in the majority of cases the two quantities have the same sign on occasion they do not. These are samples in which the skewness test would wrongly lead one to expect a degenerate solution, and we recommend the use of the more reliable criterion (11) to indicate 3PLND degeneracy instead.

While this degeneracy criterion applies strictly only to the 3PLND, we have found it to be a reliable indicator of degeneracy in the LNN as well. That is, we redefine

$$S_k = \sum_{i=1}^{n} [\ell n\, X_i + \alpha)^k] /n, \text{ and compute}$$

$$\sum_{i=1}^{n} \frac{1}{\bar{X}_i+\alpha} [\ell n\, (X_i+\alpha) - S_1 + S_2 - S_1^2] \qquad \ldots(11)$$

for a suitably large value of α, predicting degeneracy if its value is negative, and regularity otherwise. While the validity of this test is based only on the heuristic argument of analogy with the underlying 3PLND variable, to date we have found no data set in which it was wrong in either direction about convergence of maximum likelihood.

Comparison of the three estimation procedures

A simulation experiment was run to compare these three estimation procedures. Six (ξ,σ) pairs were selected and random samples of size 100 generated for each. The summary statistics of the resulting method of moments (MM) and modified method of moments (MMM) and maximum likelihood (ML) are given in Table 2.

The most striking feature of this table is the substantial bias and large variances of the estimates given by both the methods based on moments. Clearly neither method is attractive except as a preliminary way of getting starting values for maximum likelihood. It is also noteworthy that the modified method of moments fails frequently producing impossible parameter values, particularly when the logarithmic mean is large. Of the two moment methods therefore, that based on mean variance and skewness is the better, despite its apparent unattractiveness in depending on a moment with large sampling variance.

Maximum likelihood by contrast shows little evidence of noticeable bias, even though the sample size of 100 used is not of a size where one would normally invoke asymptotic consistency very confidently. It is thus to be recommended. Fortunately, convergence to the MLE's starting from the MM estimates is usually quite quick, typically occuring in 10 or fewer iterations.

4. Inference on Z for given X

The work on the LNN distribution was motivated by the framework in which there is interest in an underlying 3PLND variate which is observed with appreciable random error. A major concern in this situation is to estimate the underlying Z from the reading X. For this we require the conditional distribution of Z given X, as well as the moments of this distribution. At this point we return to the full four-parameter form of the distribution, but will work for notational convenience with $Y = Z+\alpha$ rather than Z - Z can be recovered from Y very easily by subtracting α.

The conditional distribution of $Y|X = x$ is by definition $f(x,y)/g(x|\alpha,\tau,\xi,\sigma)$

$$= \frac{1}{2\pi\sigma\tau y} \exp -\frac{1}{2}\left[\frac{(\ln y - \xi)^2}{\sigma^2} + \frac{(x - y + \alpha)^2}{\tau^2}\right] \Big/ g(x|\alpha,\tau,\xi,\sigma)$$

While this distribution is not a standard form, its moment can be found quite easily, the k-th moment being given by:

$$E[Y^k|X=x] = \int_{y=0}^{\infty} y^k f(x,y)dy \Big/ g(x|\alpha,\tau,\xi,\sigma). \quad \text{Writing}$$

$$\int_{y=0}^{\infty} y^k f(x,y)dy = \int_{y=0}^{\infty} \frac{y^k}{2\pi\sigma\tau} \exp -\frac{1}{2}\left[\left[\frac{\ln y - \xi}{\sigma}\right]^2 + \left[\frac{x + \alpha - y}{\tau}\right]^2\right] dy$$

$$= \int_{t=-\infty}^{\infty} \frac{e^{kt}}{2\pi\sigma\tau} \exp -\frac{1}{2}\left[\frac{(t-\xi)^2}{\sigma^2} + \frac{(x - e^t + \alpha)^2}{\tau^2}\right] dt$$

and incorporating the exp(kt) term into the first term in square brackets and simplifying gives for this integral

$$e^{k\xi+\frac{1}{2}k^2\sigma^2} \int_{t=-\infty}^{\infty} \frac{1}{2\pi\sigma\tau} \exp -\frac{1}{2}\left[\frac{(t-\xi-k\sigma^2)^2}{\sigma^2} + \frac{(x - e^t + \alpha)^2}{\tau^2}\right] dt.$$

The integral in this expression is recognised from the defining equations of

the LNN distribution, and so finally we may write

$$E[Y^k|X=x] = e^{k\xi+\frac{1}{2}k^2\sigma^2} \frac{g(x|\alpha,\tau,\xi+k\sigma^2,\sigma)}{g(x|\alpha,\tau,\xi,\sigma)}. \qquad \ldots(12)$$

The exp $(k\xi+\frac{1}{2}k^2\sigma^2)$ term is the marginal k-th moment of the 2PLND variate Y, so that the k-th conditional moment may be thought of as the marginal kth moment og Y multiplied by a conditioning correction factor involving the two g terms.

The moments are sufficient to solve the calibration problem. The best estimate of Y given X is E(Y|X), and the variance of prediction is $E(Y^2|X)-[E(Y|X)]^2$. As $Z=Y-\alpha$, this provides the answer to the problem of point estimation with standard error of Z from its LNN reading.

Approximating the conditional distribution of Y given X

While the mean and standard error suffice for many purposes, it is also helpful, particularly if we seek interval estimates for Z, to have the conditional distribution of Y|X. For this, consider the formal expansion

$$\ell n\, g(x|\alpha,\tau,\xi+\delta,\sigma) = \ell n\, g(x|\alpha,\tau,\xi,\sigma) + a\delta + \tfrac{1}{2}b\delta^2 + \ldots \qquad \ldots(13)$$

Then

$$\begin{aligned} \ell n\, E[Y^k|X=x] &= k\xi + \tfrac{1}{2}k^2\sigma^2 + ak\sigma^2 + \tfrac{1}{2}bk^2\sigma^4 + \ldots \\ &= k(\xi + a\sigma^2) + \tfrac{1}{2}k^2(\sigma^2 + b\sigma^4) + \ldots \qquad \ldots(14) \end{aligned}$$

This suggests that to the extent that $\ell n\, g$ can be approximated by a quadratic in σ^2, the conditional distribution of Y given X may be approximated by a 2PLND and therefore that of Z by a 3PLND. As we shall show in Section 6 by a numerical example, a lognormal approximation does indeed appear to fit well.

5. Computational details

Returning to the convenient canonical form $\alpha=0$, $\tau=1$ in terms of which both g and its derivatives may be expressed, we find that g and its derivatives involve the terms

$$A(t) = \exp -\frac{1}{2}\left[\left[\frac{t - \xi}{\sigma}\right]^2 + \left[e^t - x\right]^2\right]$$

while G involves the term

$$\exp -\frac{1}{2}\left[\frac{t - \xi}{\sigma}\right]^2 \bar{\Phi}(x - e^t)$$

The first term of each of these pairs tends to zero as $t \to \pm\infty$. The second term $\to 0$ extremely rapidly as $t\to\infty$, but tends very slowly to nonzero limit as $t\to-\infty$. Attempts were made to obtain better conditioned integrands using repeated integration by parts based on the iterated integrals of the error function (Abramowitz and Stegun 1965); but these attempts actually worsened the situation, and finally a direct evaluation using Romberg quadrature (Henrici 1963) was used. As the high-order derivatives of the integrand can vary enormously over the range of integration, we found it helpful to perform separate integrations in three subranges - a narrow central part where the high-order derivatives were large, and separate left and right tails. This provided a satisfactory algorithm with an internal estimate of accuracy and acceptable though large execution times.

Maximum likelihood estimation is the most time-consuming routine operation with LNN data since it involves many evaluations of g for each value in the sample. The initial estimates for maximum likelihood estimation are obtained from the method of moments or the modified method of moments. Where the data show little skewness (indicating that the solution though not degenerate may be quiteclose to degenerate), the initial α may be extremely large and the initial σ near zero. Not only can this produce poor convergence, but the

small σ can give computer overflow problems, and it is necessary to edit these values on occasion to produce starting values of σ sufficiently above zero to avoid these problems.

We have had good results applying maximum likelihood using a hybrid algorithm which moves smoothly between steepest-descent and Newton steps.

6. Example

As an illustration of the fitting and use of the distribution, we consider the following set of data obtained by counting a rich gold reef using the XRF portable gold analyzer (Lloyd 1980). Class midpoints (measured in centimeter-gram per tonne, or cm g/t) and frequencies were as follows:

X_i	f_i	X_i	f_i	X_i	f_i
-300	7	-100	34	100	69
300	63	500	32	700	31
900	19	1100	5	1300	12
1500	12	1700	7	1900	7
2100	2	2300	0	2500	4
2700	2	2900	3	3100	1
3300	1	3500	0	3700	0
3900	0	4100	1	4300	1
4500	1	4700	2		

In an unpublished report by H.S. Sichel, theoretical calculations using Poisson counting statistics of the instrument led to the value $\tau = 214$, a figure which was verified by actual remeasurement of some of the sampled points. The data show clearly the phenomenon of negative readings for gold, the lowest being approximately -400 cm g/t.

The modified method of moments estimation procedure fails with this data set - the estimates obtained are

$$\alpha = -407,\ e^{\xi} = -55,\ \downarrow\omega = -3.9.$$

The conventional method of moments estimator gives more usable values:

$$\alpha = 575,\ \xi = 6.9,\ \sigma^2 = 0.37.$$

Starting from the method of moments estimates, the maximum likelihood estimates, obtained in 8 iterations, are:

$$\hat{\alpha} = 46,\ \hat{\xi} = 5.929,\ \hat{\sigma}^2 = 1.217.$$

The observed Fisher information gives the following approximate (asymptotic) standard errors for the parameters and correlation matrix between estimates:-

	s.e.	Correlation		
α	35.41	1.00		
ξ	0.129	0.90	1.00	
σ^2	0.196	-0.84	-0.86	1.00

A question of some interest is how severely the measurement error impacts the quality of the parameter estimates - in other words, how much better would the estimates of the parameters be if the underlying Z were observed without error. This question can be addressed by computing the expected Fisher information from a lognormal sample of the same size and with the same parameter values - in effect mimicing what one would obtain if it were possible to measure the underlying true lognormal variables Z. This gives the following standard errors and correlation matrix:-

	s.e.	Correlation		
α	6.30	1.00		
ξ	0.069	0.44	1.00	
σ^2	0.122	-0.61	-0.27	1.00

A comparison of these two sets of figures shows that the measurement error has a large impact on the quality of the estimate of α, where the standard error is increased nearly six-fold over what would be obtained if the underlying lognormal Z were measurable directly. The effect on the other parameters is not as severe - the standard errors for the key parameters ξ and σ^2 double and increase some 50% respectively.

The correlations between the parameter estimates is high, indicating that there is a range within which one can simultaneously adjust α, ξ and σ^2 giving a series of values fitting the data about equally well. This phenomenon is well known in the underlying 3PLND distribution of Z (Johnson and Kotz 1970 p. 122), and the presence of the measurement error exacerbates it.

In application, the next question would be the calibration curve to use to estimate the grade Z from the reading X. We get this quite easily from the material of section 4, which provides expressions for E(Y|X) and var(Y|X). We were also interested in the question of how well the conditional distribution of Y given X could be approximated for each X by a 2PLND. A partial answer to this is obtained by defining the conditional logarithmic mean ξ_x and variance σ_x^2 implicitly from the equations:

$$E(Y|X=x) = \exp(\xi_x + \tfrac{1}{2}\sigma_x^2); \quad E(Y^2|X=x) = \exp[2(\xi_x+\sigma_x^2)].$$

We may then obtain an idea of the quality of the 2PLND fit by comparing the exact third and fourth moments of the conditional distribution of Y given X with the values implied by the 2PLND fit - $\exp(3\xi_x+4\tfrac{1}{2}\sigma_2^2)$ and $\exp(4\xi_x+8\sigma_x^2)$ respectively.

Table 3 shows for each class midpoint X used in the data (i) the computed conditional logarithmic mean and variance of Y given X; (ii) the natural log of the ratio of the exact third and fourth conditional moments to their 2PLND approximations; and (iii) the calibration curve producing Z and the standard error of the predicted Z. Several features of this curve deserve comment.

(i) Though the table does not show it, even when the X values are calibrated, negative predictions remain possible. In fact, as $x\to-\infty$, $E(Y|X=x)\to-\alpha$, which for this data set is -46. In reality this is a reflection, not of the quality of the LNN distribution, but on the fact that the 3PLND model used for Z is itself erroneous in permitting (albeit with low probability) negative grades.

(ii) The calibration curve is nonlinear, though as $x\to\infty$, $E(Y|X=x)\to x$. By

contrast, the regression of X on Z is a 45° line through the origin.

(iii) The conditional logarithmic variance decreases as x increases, showing that in percentage terms the estimation of Z is much more precise at higher readings than at lower. The standard error on the linear scale is low for small X but increases toward the measurement standard deviation τ as X increases. The reason for this is that where X is near or below zero, considerable use is made of the 'prior' information about the distribution of Z, but where X is large, the calibration curve approaches X itself.

(iv) Judged by the correspondence between the exact and approximate third and fourth moments, the 2PLND approximation for Y is excellent for moderate to large x and good for small x. This fact suggests that good confidence intervals for Z may be based on the approximation that for given x

$$\ell n\ (Z+\alpha) \sim N(\xi_x, \sigma_x^2).$$

7. Conclusion

In previous analyses of lognormal data, measurement error in obtaining the data has generally been ignored, except where recording error has been used as a computational device to avoid singularity problems. Increasingly, use is being made of instrumental methods, or indirect assays, where there is substantial measurement error. As these measurements may be far quicker or more economical than direct measurement of the underlying lognormal quantity, they provide a favorable tradeoff of more observations at less cost (see for example Magri and Hawkins 1987), and so their use is likely to increase in future.

We have derived expressions for the lognormal - normal convolution that arises when the measurement is unbiased and of constant variance. Also given are formulas for the quantities necessary for most applications - estimation, calibration and tabulation.

References

Abramowitz, Milton and Stegun, Irene A. (1965), Handbook of Mathematical Functions, New York: Denver.

Aitchison, John and Brown J.A.C. (1957), The Lognormal Distribution, London: Cambridge University Press.

Cohen, A. C., and Whitten, B. J., (1980), 'Estimation in the three-parameter lognormal distribution', Journal of the American Statistical Association, 75, 399 - 404.

Griffiths, David A. (1980), 'Interval Estimation for the Three-Parameter Lognormal Distribution via the Likelihood Function,' Applied Statistics, 29, No. 1, 58-68.

Harbaugh, John W. and Ducastaing, Michel (1981), 'Historical Changes in Oil-Field Populations as Method of Forecasting Field Sizes of Undiscovered Populations: A Comparison of Kansas, Wyoming and California,' Subsurface Geology Series 5, Lawrence: Kansas Geological Survey.

Harter, H. Leon, and Moore, Albert H., (1966), 'Local maximum likelihood estimation of the parameters of three-parameter lognormal populations from complete and censored samples', Journal of the American Statistical Association, 61, 842 - 851.

Henrici, Peter (1963), Elements of Numerical Analysis, New York: Wiley.

Johnson, Norman L. and Kotz Samuel (1970), Continuous Univariate Distributions - 1, Boston: Houghton Mifflin.

Klimko, L. A., Rademaker, A., and Antle, C. E., (1975), Communications in Statistics, 4, 1009-1019.

Kotz, S., (1973), Communications in Statistics, 1, 113-132.

Krige, Daniel G. (1960), 'On the Departure of Ore Value Distributions from the Lognormal Model in South African Gold Mines,' Journal of the South African Institute for Mining and Metallurgy, 61, 231-244.

Krige, Daniel G. (1978), Lognormal - De Wijsian Geostatistics for Ore Valuation, South African Insititute for Mining and Metallurgy, Johannesburg.

Lloyd, Philip J.D. (1980), 'Portable Gold Analyser for the Mines,' South African Journal of Science, 76, 440-441.

Magri, E., and Hawkins, D. M., (1987), 'Possible improvements in mine valuation due to the introduction of the gamma ray fluorescence gold ore analyser', Journal of the South African Institute for Mining and Metallurgy, 87, 407-423.

Matheron, Georges (1960), Treatise on Applied Gestatistics, Kennecott Copper Corporation, Salt Lake City.

Rolle, Rainer (1979), Gamma Ray Fluorescence for in situ Evaluation of Ore in Witwatersrand Gold Mines, Ph.D. Thesis, University of the Witwatersrand, Johannesburg.

Table 1. The median of the canonical LNN distribution for different ξ and σ

$\sigma\backslash\xi$	-2	-1	0	1	2	3	4	5	6	7	8	9	10
0.1	1.00	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.2	1.02	1.02	1.02	1.02	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.3	1.05	1.05	1.04	1.03	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.4	1.08	1.08	1.07	1.04	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.5	1.13	1.13	1.10	1.04	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.6	1.19	1.18	1.13	1.05	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.7	1.27	1.24	1.16	1.05	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.8	1.36	1.30	1.18	1.06	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.9	1.46	1.37	1.21	1.06	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1.0	1.57	1.43	1.23	1.06	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1.1	1.69	1.50	1.25	1.06	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1.2	1.82	1.56	1.27	1.07	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1.3	1.95	1.63	1.29	1.07	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1.4	2.09	1.69	1.31	1.07	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1.5	2.23	1.74	1.32	1.07	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1.6	2.36	1.80	1.34	1.07	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1.7	2.50	1.86	1.35	1.07	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1.8	2.63	1.91	1.36	1.07	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1.9	2.77	1.96	1.38	1.07	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2.0	2.90	2.01	1.39	1.07	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 2

ξ	σ^2		α		ξ		σ^2		Number
			Mean	sd	Mean	sd	Mean	sd	successful
0	1	MMM	2.89	2.01	1.33	0.43	0.22	0.12	45
		MM	2.15	1.09	1.16	0.35	0.28	0.21	50
		ML	0.25	0.57	0.16	0.53	0.93	0.57	50
0	4	MMM	59.77	119.19	3.45	1.03	0.34	0.18	35
		MM	12.24	12.98	2.31	0.54	0.90	0.23	50
		ML	0.00	0.27	-0.14	0.49	4.28	1.59	50
5	1	MMM	131.18	134.88	5.60	0.37	0.54	0.19	14
		MM	123.70	61.79	5.64	0.27	0.52	0.21	50
		ML	-5.05	7.07	4.96	0.12	1.12	0.25	50
5	4	MMM	8914	14104	8.58	0.91	0.39	0.19	30
		MM	1779	1552	7.34	0.46	0.93	0.23	50
		ML	-0.69	1.34	4.98	0.24	4.25	0.60	50
-1	1	MMM	3.35	3.59	1.06	0.74	0.10	0.10	30
		MM	21.73	86.21	1.52	1.21	0.11	0.21	50
		ML	4.38	7.30	-0.05	1.92	1.00	1.36	50
-1	4	MMM	15.64	17.24	2.35	0.93	0.31	0.17	48
		MM	4.16	2.59	1.35	0.41	0.87	0.25	50
		ML	0.05	0.21	-0.95	0.62	3.95	1.68	50

Table 3

x	ξ_x	σ_x^2	Approximation to μ_3'	Approximation to μ_4'	E(Z|X)	se(Z|X)
-300.	4.6131	0.3523	-0.094	-0.344	74.58	78.12
-100.	4.8650	0.3444	-0.101	-0.366	108.41	98.77
100.	5.1876	0.3172	-0.103	-0.366	164.17	128.18
300.	5.5871	0.2598	-0.089	-0.313	258.37	165.58
500.	6.0242	0.1789	-0.058	-0.202	406.38	200.08
700.	6.4158	0.1072	-0.027	-0.097	599.47	217.03
900.	6.7186	0.0640	-0.011	-0.041	808.97	219.71
1100.	6.9497	0.0414	-0.005	-0.018	1019.02	218.79
1300.	7.1345	0.0289	-0.002	-0.009	1227.10	217.80
1500.	7.2886	0.0213	-0.001	-0.005	1433.56	217.07
1700.	7.4210	0.0164	-0.001	-0.003	1638.83	216.56
1900.	7.5372	0.0130	-0.001	-0.002	1843.22	216.14
2100.	7.6409	0.0106	-0.000	-0.001	2046.94	215.86
2300.	7.7344	0.0088	-0.000	-0.001	2250.15	215.59
2500.	7.8198	0.0074	-0.000	-0.001	2452.96	215.29
2700.	7.8982	0.0063	-0.000	-0.000	2655.41	215.16
2900.	7.9708	0.0055	-0.000	-0.000	2857.60	215.09
3100.	8.0384	0.0048	-0.000	-0.000	3059.55	214.98
3300.	8.1017	0.0042	-0.000	-0.000	3261.31	214.88
3500.	8.1611	0.0037	-0.000	-0.000	3462.90	214.82
3700.	8.2171	0.0033	-0.000	-0.000	3664.34	214.89
3900.	8.2701	0.0030	-0.000	-0.000	3865.70	214.59
4100.	8.3204	0.0027	-0.000	-0.000	4066.90	214.79
4300.	8.3683	0.0025	-0.000	-0.000	4268.05	214.53
4500.	8.4140	0.0023	-0.000	-0.000	4469.09	214.53
4700.	8.4576	0.0021	-0.000	-0.000	4670.05	214.71